NLP-Based Fake News Classifier

Project for E4990 Introduction to Data Science Industry

Team Members:

Jason Lei and Andres Soto

Roles:

Machine learning and Web development

Dataset:

https://www.kaggle.com/mrisdal/fake-news

Project Description and Motivation

While reading about available open-source datasets, we discovered the Fake News Dataset, which contains text and metadata from fake and biased news sources on the web. This was immediately interesting to us, as handling fake news is becoming an increasingly divisive topic. There are claims that fake news influenced our most recent presidential election and, generally, misinformation helps no one. With the rise of social media, news and information are being spread more and more by people, instead of through established channels like television or newspaper. Thus, with this notion of virality, people need to be able to easily distinguish between real and fake news to make informed decisions. This project aims to provide a channel for users to check the likelihood of their news source being fake; we aim to create a classifier that will predict whether a given input is real or fake news, based on a lexical-and-syntactical-analysis approach.

Audience

Anyone who reads online news sources and is unsure about whether their news source is reliable or not.

Algorithms

This is a supervised classification problem, as we have labels for all of our training data (although the objectivity of this labeling can always be debated). We will use Logistic Regression or Support Vector Machines. We will start by using logistic regression to establish a baseline classification model for whether or not a block of input text is fake news or real news. Then, as we find more features, we may try an SVM approach. (SVMs are supposed to perform better in spaces with a high number of dimensions)

Some difficulties I imagine facing will include: completing the dataset by scraping the web for non-fake news sources to compare our current dataset to, obtaining a lexical and syntactical breakdown of our dataset, and ranking the most important features for our model. (e.g. Negative words? Use of adjectives? Punctuation? Capitalization?)

Interface

From my broad proposal standpoint, I think that a lot of the challenges in this project will come from the web application itself, which I don't have much experience building. I'd like to create a web app that allows users to input a block of text from their news source and then our application will classify the text based on the model we've created. I imagine it will be difficult to make our model available through the web interface. Once this is completed, the next step would be to add web-scraping abilities to our web application, so that users can simply input a web link and then the application will obtain the text on its own.